Numeric Data

Citation Techniques and Integration with Text

Brasen, Jan; Farquhar, Adam; Gastl, Angela; Gruttemeier, Herbert; Heijne, Maria; Heller, Alfred; Hitson, Brian; Johnson, Lorrie; McMahon, Brian; Piguet, Arlette

Total number of authors:
13

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Numeric Data: Citation Techniques and Integration with Text

Jan Brase[1], Adam Farquhar[2], Angela Gastl[3], Herbert Gruttemeier[4], Maria Heijne[5], Alfred Heller[6] , Brian Hitson[7], Lorrie Johnson[7] , Brian McMahon[8] , Arlette Piguet[3], Jeroen Rombouts[5], Mogens Sandfaer[6] and Irina Sens[1],

1. German National Library of Science and Technology (TIB)
2. The British Library
3. ETH Library Zürich
4. Institute for Scientific and Technical Information (INIST) -CNRS
5. TU Delft Library
6. Technical Information Center of Denmark
7. U.S. Department of Energy/ Office of Scientific and Technical Information
8. International Union of Crystallography

Disclaimer

This report is the outcome of the project “Numeric Data: Citation Techniques and Integration with Text” from the *Technical Activities* Coordinating Committee (TACC) of the International Council for Scientific and Technical Information (ICSTI).

Parts of this report are also published as

“Approach for a joint global registration agency for research data”
Jan Brase, Adam Farquhar, Angela Gastl, Herbert Gruttemeier, Maria Heijne, Alfred Heller et al
Information Services & Use (2009) 1–15
doi:10.3233/ISU-2009-0595
IOS Press

Table of contents

Abstract

The scientific and information communities have largely mastered the presentation of, and linkages between, text-based electronic information by assigning persistent identifiers to give scientific literature unique identities and accessibility. Knowledge, as published through scientific literature, is often the last step in a process originating from scientific research data. Today scientists are using simulation, observational, and experimentation techniques that yield massive quantities of research data.

These data are analysed, synthesised, interpreted, and the outcome of this process is generally published as a scientific article. Access to the original data as the foundation of knowledge has become an important issue throughout the world and different projects have started to find solutions.

Global collaboration and scientific advances could be accelerated through broader access to scientific research data. In other words, data access could be revolutionized through the same technologies used to make textual literature accessible.

The most obvious opportunity to broaden visibility of and access to research data is to integrate its access into the medium where it is most often cited: electronic textual information. Besides this opportunity, it is important, irrespective of where they are cited, for research data to have an internet identity.

1. Background

Knowledge, as published through scientific literature, often is the last step in a process originating from research data. These data are analysed, synthesised, interpreted, and the outcome of this process is generally published in its result as a scientific article.

Only a very small proportion of the original data are published in conventional scientific journals. Existing policies on data archiving notwithstanding, in today's practice data are primarily stored in private files, not in secure institutional repositories, and effectively are lost [LAW01]. This lack of access to scientific data is an obstacle to international research. It causes unnecessary duplication of research efforts, and the verification of results becomes difficult, if not impossible ([DIT01]). Large amounts of research funds are spent every year to re-create already existing data ([ARZ04]).

Data have always been at the heart of scientific progress. They are the raw material out of which research can be carried out and what many publications are based upon. Data integration with text is therefore an important aspect of scientific collaboration. It allows verification of scientific results and joint research activities on various aspects of the same problem. Data integration is instrumental for the successful realization of multidisciplinary research, academia-industry collaboration and the development of new products in large scale engineering projects (e.g. in the aerospace, ship building or automotive industries).

Recognizing the need for data sharing, several scientific communities have organized data collection, archiving and access to serve their community needs. For instance, earth and environmental studies data are collected and shared on a world-wide level through the World Data Center System (http://www.ngdc.noaa.gov/wdc/). Data publication is an essential component of every large scientific instrument project (e.g., the CERN Large Hadron Collider). In fact, the development of grid technology can be linked to infrastructure requirements that were raised by the volume of information that high energy physics experiments generates and by the need to share this information among physicists across the globe. Similar examples can be found in geophysics, chemistry, astronomy, biology, etc.

Progress in sharing of scientific data has been made at a fast pace. Infrastructures such as grid exist for storage. Methodologies have been established by data curation specialists to build high quality collections of datasets. These include standards for metadata (provenance, copyright, author of a dataset), registration, cataloguing, archiving and preservation. A large number of disciplines benefit from these methodologies and high quality datasets. **Fehler! Verweisquelle konnte nicht gefunden werden.**illustrates how formal dataset publication effectively transforms data into information and ultimately knowledge.

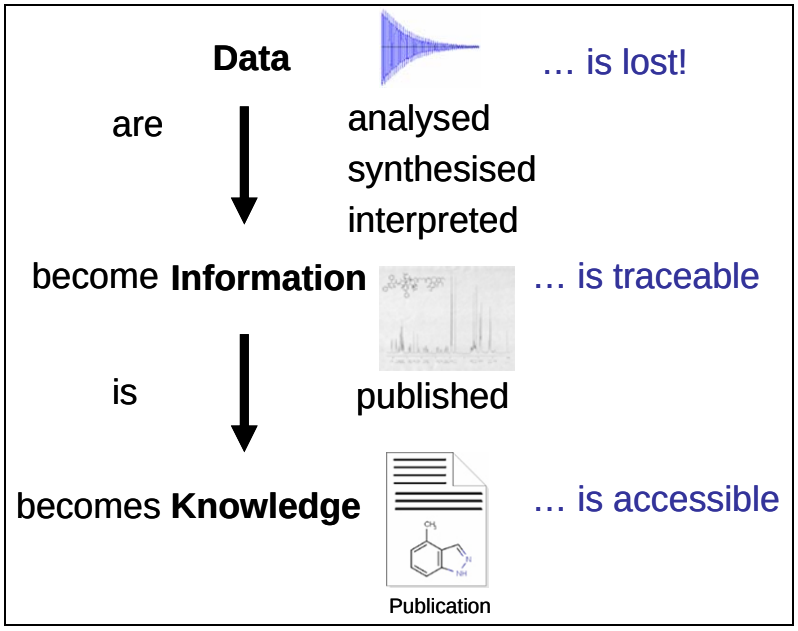

Figure 1 - From Data to Knowledge through publication

1.2 Issues

Unfortunately, a large body of data used for research is not published following established best practices.

Problem 1. A large volume of research data is not shared at all. Since academic recognition is mainly achieved through publication, sharing datasets is a time consuming task not adequately compensated. In addition, other considerations such as the researcher liability in releasing datasets, unclear dataset ownership, or the unavailability of a repository to the researcher are factors that hinder data sharing best practices;

Problem 2. When published, datasets often do not follow the same process as articles. While articles are duly incorporated in digital libraries and can be referenced

– in a persistent manner – in other articles, datasets are not published, or published only on the researcher's web site and, if referenced at all, only referenced by the corresponding URL. Such publication model raises a number of issues (see Figure 2):

(i) Poor preservation properties (e.g. if the researcher moves to another institution, the link may become invalid);

(ii) Poor quality of the documentation;

(iii) Limited impact and academic recognition (dataset cannot be searched or found except from article reference or web search);

(iv) Lack of data quality assessment.

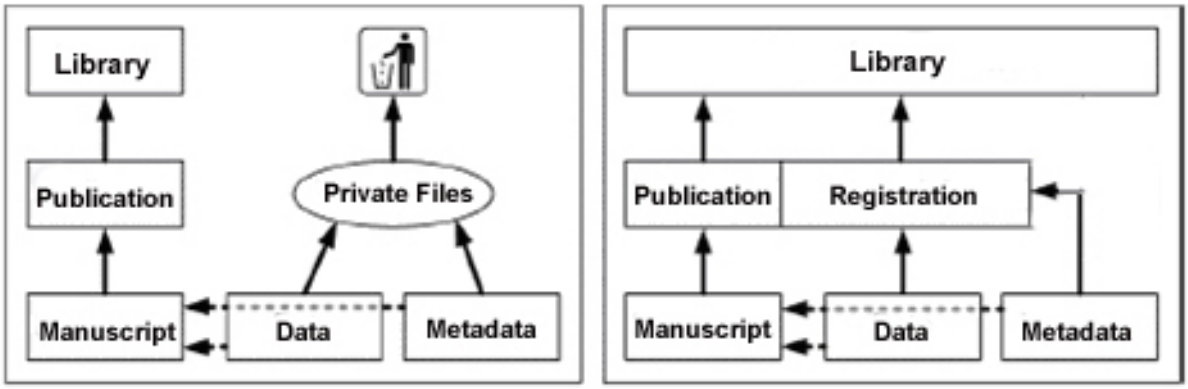

Figure 2-The traditional publication method for datasets on the left, a possible new structure on the right

2. Data Citation Techniques and Integration with Text

Integration of data into texts and unlocks citation services provides incentives for a researcher to publish datasets. Moreover, the publication of datasets and the inclusion of the dataset into library catalogues improves its potential impact since more researchers will become aware of its availability.

Currently, a large number of datasets are either directly referenced through their location (e.g., a URL) or through community-specific registries. URLs are subject of preservation issues since location of resources may change over time. Community specific registries introduce interoperability issues (due to heterogeneity in resolution services). Moreover, resolution of such identifier is difficult, since often one has to first identify the specific resolution engine that was used for issuing the identity.

In a number of scientific communities, there is no established data repository or dataset quality assessment protocol. In such cases, datasets are published by

researchers using ad-hoc approaches. For instance, a group may make a dataset available on a web-page and communicate about it through an article, providing a URL to the dataset. Such a publication model raises preservation and quality concerns. Documentation of the dataset, if provided, would typically be based on an ad-hoc document format. License and copyright for using the dataset would often be unclear and it would not be possible to leverage metadata harvesting protocols for improving the visibility of the dataset. Publishing a dataset using standards compliant methodology (e.g., with Dublin-core metadata) is time consuming. In addition, these protocols are often not known by researchers, as dataset curation is not part of their specialties and tasks.

For academic researchers, dataset publication is not rewarded by an academic recognition proportionate to the effort. For a dataset to “count” as a publication, it would need to follow similar publication process as an article: be properly documented, be reviewed for quality, be searchable in catalogues, and be citable in articles. Moreover, in a way equivalent to citation count for articles, dataset usage needs to be measured to provide an indication of impact on the scientific community and a driver of academic recognition.

Research is a global endeavour and dataset identification and cross-referencing shall be accomplished at a global level. Raising the awareness of researchers of available datasets is also important for providing the best research possible. Similarly, publicizing the availability of dataset resources worldwide is essential, to achieve a full valorization.

2.1 Global awareness

Access to research has become an important issue throughout the world, identified by different organisations and individuals.

In its 2007 report “Cyberinfrastructure Vision for 21st Century Discovery“ [NSF07] the National Science foundation (NSF) remarks:

“*Science and engineering research and education have become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared and analyzed.*

Worldwide, scientists and engineers are producing, accessing, analyzing, integrating and storing terabytes of digital data daily through experimentation, observation and simulation. Moreover, the dynamic integration of data generated through observation and simulation is enabling the

development of new scientific methods that adapt intelligently to evolving conditions to reveal new understanding. The enormous growth in the availability and utility of scientific data is increasing scholarly research productivity, accelerating the transformation of research outcomes into products and services, and enhancing the effectiveness of learning across the spectrum of human endeavour"

In 2007 the Organisation for Economic Co-operation and Development (OECD) has published their "OECD Principles and Guidelines for Access to Research Data from Public Funding" [OECD07]. It identifies the important aspects from the perspective of the public funders:

"The rapid development in computing technology and the Internet have opened up new applications for the basic sources *of research — the base material of* research data *— which has given a major impetus to scientific work in recent years. […].*
Besides, access to research data increases the returns from public investment in this area; reinforces open scientific inquiry; encourages diversity of studies and opinion; promotes new areas of work and enables the exploration of topics not envisioned by the initial investigators."

Further initiatives or reports addressing the issue of research data are for example:

- *The Interagency Working Group on Digital Data* (IWGDD) in the U.S. [NAT07]
- The report *Shared Responsibilities in Sharing Research Data: Policies and Partnerships* by the European Science Foundation (ESF) and the German Research Foundation (DFG) [ESF08].
- The *Strategic Coordinating Committee on Data and Information* (SCCID) established by the International Council for Science (ICSU).
- The *Digital Curation Center* (DCC) in the UK (http://www.dcc.ac.uk/about)
- The European *Alliance for Permanent Access* (http://www.alliancepermanentaccess.eu/index.php?id=1).
- The *UK Research Data Service* (UKRDS) (http://www.ukrds.ac.uk/)
- *Australian National Data Service* (ANDS) [AUS07]
- *Research Data Canada*, by the National Consultation on Access to Scientific Research Data (NCASRD), Canada (http://data-donnees.gc.ca/eng/ncasrd/index.html)

2.2 State-of-the-art: Data infrastructures in Europe

Europe has a large number of infrastructures that cater for dataset publication needs at various levels.

Horizontal infrastructures are providing generic ICT services for dataset publication, storage or processing. The pan European backbone network GEANT and the Enabling Grids for E-sciencE (EGEE) project are representatives of such horizontal infrastructures, providing respectively connectivity and grid services. They form the bottom layer of commodity services (data storage, data transport, computation, etc.) that may be used for any sort of research, from physics to biology through social science.

In contrast, vertical infrastructures provide community specific solutions for achieving data sharing in a particular discipline. These solutions typically cater for all the steps in the dataset publication workflow, allowing online submission of datasets, their registration, their publication as well as advanced search and exploration using graphical user interfaces. Examples include:

- Art and humanities repositories: Archeology Data Service (ADS) TextGrid; Netherlands Historical Data Archive (DANS);
- High energy physics: CERN, DESY;
- Biology: BioGRID (interaction dataset);

Ideally, vertical infrastructures for data curation would be built on top of lower level infrastructures. This is indeed the case. For instance, any digital library or repository today leverages the ubiquitous connectivity offered by the Internet. However, higher level functionalities are largely based on in-house developments that are not easily interoperable between communities. This is easily explained considering that:

1. **Legacy:** At the time of setting up community infrastructure, no high level commodity for dataset curation was available. This has forced communities to implement their own solutions to similar problems. Due to national initiatives, it is also common to find several distinct vertical solutions addressing a single scientific community.
2. **Heterogeneity:** Research datasets are very heterogeneous in form, complexity, size and nature. Radically different requirements from discipline to

discipline make a one-size-fit-all approach to vertical solutions doomed to failure. Integration and homogeneity are desirable but should not be achieved at the expense of truly functional solutions closely aligned with the needs of the community that use it. Therefore, generic high level solutions would require extensive customisation to address community-specific requirements.

Art and Humanities

DARIAH is an FP7 funded project that started in September 2008. It will provide research infrastructure for digital research and preservation and aims at bringing information users, information managers and information providers within countries and across Europe together. The vision for DARIAH is to facilitate long-term access to, and use of all European arts and humanities and cultural heritage information for the purposes of research.

Systems Biology

SIDR stands for “Standards-based Infrastructure with Distributed Resources” and is a French project concerned with interoperability among system biology repositories. In systems biology, research communities that produce resources face major issues related to data sharing and exchange. In particular, they need to use agreed-upon standards, ontologies, controlled vocabularies, exchange languages, etc. for the annotation of resources. The creation of an infrastructure for distributed resources, which would be the outcome of SIDR, aims to address the above issues by providing research teams with a well-structured access point to shared resources whose quality will be specified and guaranteed according to international domain standards. The CNRS' Standard-based Infrastructure with Distributed Resources (SIDR) initiative aims at building a resource centre with international dimension that will enable dissemination, adding value and sharing of resources, including quantitative data in the systems biology.

Finance and economics

SIRCA is a not-for-profit financial services research organization involving twenty-six collaborating universities across Australia and New Zealand. SIRCA is managing large repository of datasets. For instance, SIRCA is managing the archives of Reuter’s financial news, corresponding to 10 years of news and stock market

movement and totalizing over 100 terabytes. This dataset forms a single very large set of records utilizing an ad-hoc representation. In this archive, all news or stock events are represented as a single stream of information, corresponding to the method by which this type of data is broadcasted to Reuter's customer terminals.

A difficulty in this context is to identify and reference relevant subsets of the archive that may be used by researchers for a given study. For instance, a researcher might extract from the archive all the news relating to Microsoft, Google and Yahoo over the last 4 years and study, say, correlation among the news. When this researcher publishes her findings, she faces the problem of how to reference the dataset that was used to derive her results. Additionally, since Reuter's financial data are copyrighted, researchers are not entitled to redistribute them. Other researchers may access the data but would have to do so through SIRCA's portal with conditions that depend on their affiliation.

Environmental air pollution monitoring

GENESIS is an FP7 project focused on environmental and air pollution monitoring at a European level. One of the problems addressed in GENESIS concerns the largely non-interoperable sources of data that are collected by different member states using each distinct metadata structure.

Without high quality metadata, air pollution survey or other environmental measures are useless since they cannot be interpreted correctly. Examples of metadata that are obligatory for interpreting such dataset include the exact location where measures where performed, which sampling methodology was used and what was the date and time at which measures where carried out. Currently, these metadata may be totally absent from data records. This is the case when data are on some file while information on the provenance is in, e.g., a report. Or, when present, metadata may use very different methodology for relating similar information (e.g., localization).

When computing environmental models at European level, metadata interoperability is a preliminary to any data integration.

Automotive Industry

Simulations in the automotive industry produce large amounts of data, created in cross-organisational workflows where typically one automotive company and several suppliers work together for a certain time in a project while, at the same time, they

may potentially be competitors in another context. An important requirement for improving collaboration and projects based on shared data is the management of large data sets that are in the size of several tens to hundreds Gigabytes, the ability to cope with lots of different non-standardised formats, as well as protection of IPR and data privacy during the entire workflow. Legal reasons (product liability) require taking into account data provenance and metadata.

Data in automotive industry is characterised by a large heterogeneity. The steps of a workflow consider different types of data, ranging from CAD data, CAE model, parameter files, and material data in form of tables or formulae, property data and up to CAE simulation results. Additionally, even within the same type of data sets, the formats vary since the formats are typically proprietary (ISV and application specific).

Examples of dataset managed in automotive industry include:

Experimental data, on the level of material properties as well as measurements of e.g. flow behaviour or crash tests;

Data is created in the design process of a component in the form of a CAD file;

On basis of CAD files, the simulation models for CFD (Computational Fluid Dynamics) or CSM (Computational Structural Mechanics) are created;

Input parameter files for the CAE software tools as well as scripts are prepared on the basis of material and properties data and simulation code parameters;

Simulation result data

This indicates the pressing need for a coherent data management in the context of automotive collaborative project. A number of constraints make collaborative data management more difficult:

Privacy requirements: it varies from publicly released results in publications to closed, OEM product-related internal data. Typically, parts of the data are shared during a project, where the important point is to track the context that lead to a shared data set. Some data might be published after a given time.

Absence of shared database systems: Typically, data are owned and hosted by one specific entity. On request, subsets of the data will be extracted and are transferred to the project partner by hand, e.g. attached to e-mails. Creation of particular subsets might be manual or (semi)automated. In automotive industry, partners are usually not

allowed to search in other partners' databases, but they need datasets for given purpose. Dataset identification is then defined individually according to the providing partners' format. While the automotive companies (OEMs) typically have defined identification systems, especially small and medium enterprises (SME) suppliers and engineering consultants often have not defined such general processes.

Complexity of the data processing: Automotive simulation is a multifold and often non-linear process. The first step is the design idea, followed by construction in CAD which might lead to large single files containing all parts and connectors of a device. For CAE simulations, specific areas have to be extracted, i.e. sub-files are created out of the master file. The other way round is also usual: different parts are constructed individually leading to several files that have to be joint to an overall construction of the entire device.

Data format heterogeneity: Most CAD tools use their own proprietary file format. Thus, data exchange between different CAD systems is difficult. Attempts for standardisation exist, mostly DXF format for drawings. Nevertheless, most systems read and write DXF only as 2D data, system-specific information get lost or cannot be represented appropriately in the other system. System neutral data formats are VDA-FS, IGES, and STEP and for special applications the STL data format.

VDA-FS – Data exchange format for surfaces, developed by Verein Deutscher Automobilindustrie (VDA), in the past a quasi-standard in automotive CAD;

IGES – Data exchange format for 2D-drawings and 3D surfaces, in most CAD applications implemented. More flexible than VDA-FS, more comprehensive and system-independent than DXF;

STEP – standardised data exchange format, international development aiming at exchanging parameterised data. Exchange of solid and volume data nearly loss-free and with parametric (in solids).

Documentation of data provenance: While provenance is partially done, there is no standardised format or data set that must be recorded. Furthermore, the traceability in particular if several data processing steps are involved is not yet realised. So far some of the steps are recorded but the provenance for a cross organisational workflow spanning several potentially long lasting steps or as part of a large parameter study is not implemented. In particular, the collection of provenance data is largely a manual process not properly supported by automation.

Data publication and reference management: the link between available simulation data and/or other kind of data sets such as CAD files to scientific publications is only indirectly possible via the authors of papers or reports. Additionally, the publication of data, even within a controlled group of consumers cannot be easily related to publications as quite typically the collaborators do not operate on the full data sets but on temporary “snapshots” of a full data set that only contain the information necessary to perform their respective tasks. The lifetime of such snapshots is limited to the processing task and not foreseen for long term storage.

Aerospace industry

A key technology in aerospace research and development is high-resolution parallel simulation on supercomputers using sophisticated numerical algorithms and optimized codes. Examples for current large-scale simulations are:

The complete simulation of all flow phenomena throughout the entire flight envelope including the multidisciplinary simulation of all involved disciplines of space and aerospace vehicles.

The multidisciplinary optimization of the overall aircraft design as well as the design of major parts, such as the turbine engines.

The goals are to analyze the aerodynamic and aero elastic behaviour of the aircraft and its parts and the numerical prediction of aircraft performance and handling qualities prior to the first flight. Similar goals are also apply to other industrial sectors, for example ship building. For these kinds of complex simulations, two distinct technologies are need. First, highly sophisticated and optimized numerical simulation codes for each involved discipline (for example, codes for computational fluid dynamics, structural analysis, or flight mechanics). Secondly, an efficient simulation infrastructure and well-designed supporting tools.

An example for a simulation and data management infrastructure is developed in the ongoing national project AeroGrid, where a Grid-based environment for collaboration between industry, research labs, and academia in the field of turbine engine design is build. AeroGrid currently addresses basic Grid-technology questions, such as deploying a Grid infrastructure, Grid-enabling applications, and defining business and security concepts.

2.3 State-of-the-art in the U.S. Department of Energy

Over the past few decades, research conducted by the U.S. Department of Energy (DOE) and its contractor organizations have created a wealth of scientific and technical data. Much of this data is housed in a variety of distributed repositories, and new data sets are continually being added at a rapid pace. Due to the dispersed nature of the data repositories, accessing the numeric data sets themselves is a genuine challenge for most potential users, particularly those who are new to a field or looking for experimental or observational data outside their normal field of expertise.

The Office of Scientific and Technical Information (OSTI), part of DOE's Office of Science, created an inventory of DOE data repositories in 2008. This information tool, called the DOE Data Explorer (DDE), includes a database of citations to over 275 data-hosting websites within the DOE complex. Along with descriptions of each data repository, links are provided to the websites, which reside at national laboratories, data centers, user facilities, colleges and universities, and other organizations funded either in whole or in part by DOE. Each individual data repository offers its own method/interface for accessing the data it houses. Some repositories provide very specialized interfaces, allowing users to search data, compare and visualize data sets, and package data for download and reuse. Other data repositories simply store the raw data files, which makes it very difficult for users to find and use this data for future research.

In addition to the many smaller data repositories cataloged by the DDE, there are nine major data centers funded by DOE listed below:

Alternative Fuels and Advanced Vehicles Data Center (AFDC): This online center, funded by DOE's Office of Energy Efficiency and Renewable Energy, is a collection of information on alternative fuels and the vehicles that use them. Alternative fuels described here are those defined by the Energy Policy Act of 1992, including biodiesel, electricity, ethanol, hydrogen, natural gas, and propane.

Atmospheric Radiation Measurement (ARM) Data Centers: ARM is a multi-laboratory, interagency program for improved scientific understanding of the fundamental physics related to interactions between clouds and radiative feedback processes in the atmosphere. ARM focuses on obtaining continuous field

measurements and providing data products that promote the advancement of climate models. The Office of Science funds this suite of data centers with locations and/or data storage at Brookhaven National Laboratory (BNL), Oak Ridge National Laboratory (ORNL), and the Pacific Northwest National Laboratory (PNNL).

Carbon Dioxide Information Analysis Center (CDIAC): CDIAC, which includes the World Data Center for Atmospheric Trace Gases, has served as the primary climate-change data and information analysis center for DOE since 1982. The Office of Science funds CDIAC, which is located at the Oak Ridge National Laboratory (ORNL) in Oak Ridge, Tennessee.

Comprehensive Epidemiological Data Resource (CEDR): CEDR is a DOE public-use repository of data from occupational and environmental health studies of workers at DOE facilities and of nearby community residents. In 1990, the Department of Health and Human Services assumed responsibility for many aspects of the epidemiology program and provides data to CEDR. The Office of Health, Safety, and Security funds CEDR, which is maintained at Lawrence Berkeley National Laboratory (LBNL) in Berkeley, California.

Controlled Fusion Atomic Data Center (CFADC): CFADC's mission is to compile, evaluate, recommend, and disseminate atomic and molecular collision data relevant to fusion energy research and development. Under different names, it has been a data center since 1958. CFADC, located at Oak Ridge National Laboratory (ORNL) in Oak Ridge, Tennessee, is funded by the Office of Science.

DOE Joint Genome Institute's (JGI) Genome Web Portal: The JGI makes high-quality genome sequencing data freely available to the greater scientific community through its web portal. Having played a significant role in the federally funded Human Genome Project--generating the complete sequences of Chromosomes 5, 16, and 19--the JGI has now moved on to contributing in other critical areas of genomics research. Funded by the Office of Science, the JGI's Genome Web Portal is maintained at the JGI Production Genomics Facility in Walnut Creek, California.

National Nuclear Data Center (NNDC): The NNDC collects, evaluates, and disseminates nuclear physics data for basic nuclear research and for applied nuclear technologies. Information available comes from the combined efforts of the NNDC, cooperating data centers, and other U.S. and international groups. The Office of

Science is the primary funding source for the NNDC at Brookhaven National Laboratory (BNL), Upton, New York.

Renewable Resource Data Center (RReDC): The RReDC provides information on several types of renewable energy resources in the United States in the form of publications, data, and maps. An extensive dictionary of renewable energy related terms is also provided. The RReDC is funded by the Office of Energy Efficiency and Renewable Energy. It is maintained at the National Renewable Energy Laboratory.

U.S. Transuranium and Uranium Registries (USTUR): The DOE-funded USTUR is operated by Washington State University in Richland, Washington. Its main product is data and information about the intake, deposition, translocation, retention, and dosimetry of the uranium, plutonium, americium, and thorium (actinide elements) in the human body. Information about the health effects of these radioactive elements in the human body is an additional product.

Historically, the DOE data centers and repositories have operated independently. Thus, each data center relies primarily on their own software toolkits for searching and accessing the data sets they maintain. In order to make the data more accessible and available to potential users, DOE/OSTI funded a Small Business Technology Transfer (STTR) grant to investigate the feasibility of assigning DOI's to specific data sets. The STTR project, begun in late 2008, is working with the ARM Data Center to assign DOI's to selected data sets using the Data Registration Agency at TIB, the German National Library of Science and Technology (see Section 3.4). Once DOI's have been assigned to particular data sets, a prototype system will be built to demonstrate the linkages between scientific publications and the actual numeric research data stored at the ARM Data Center.

The specific outcome of this project will be to highlight the "before" and "after" differences in access to numeric data sets which are commonly cited in scientific literature. While such access is beginning to become more common with respect to journal literature, the ability to link to source data from non-conventional literature is quite uncommon. This project will serve as a proof of concept for the value of assigning DOIs to data sets cited in technical reports, in particular, and will also provide instructive processes for routinizing this practice into as common a task as adding footnotes and key metadata.

3. Dataset registration

Dataset identification is a key element for allowing citation and long term integration of datasets into text as well as supporting a variety of data management activities. Also, to foster a culture of data integration, scientists need to be convinced that preparing their data for online publication is a worthwhile effort. It would be an incentive to the author if a data publication had the rank of a citeable publication, adding to his reputation and ranking among his peers. To achieve the rank of a publication, a data publication needs to meet the two main criteria, persistence and quality. Whereas the latter is a very difficult concept that should be made part of the workflow of data integration in the data producers, data persistency is a rather simple problem.

Simply making data available on the ‘web’ is not sufficient. The location of internet resources, and thus their URL, may easily change, which in most cases means to the user that data are lost ([KOE04]). This happens, for instance, if the data are deposited by a researcher in his personal page and the researcher moves from one institution to another. Additionally, this method of data publication makes very little impact since the way by which the dataset may be discovered by another researcher is either:

- Through a web search: Although scientific publications can easily be found through a web search, using the title as a stabile metadata element, the lack of well-defined titles and other metadata makes web-search for datasets difficult. The probability of a page containing the dataset to be found will mainly depend on the quality of the description that surrounds it on the page.
- Through the information in an article: Sometimes the information in an article enables readers to actually identify the location of a dataset, or at least provide contact information of the researcher who collected the data.

Both methods of accessing the dataset have clear limitations in terms of the potential impact of a dataset. It is not surprising that researchers naturally tend to focus their efforts on article publication instead of dataset publication.

For encouraging dataset publication, both the identification of dataset and the awareness of researcher of the availability of this dataset have to be dramatically improved.

3.1 Identifier Schemes

Identification of electronic resources through persistent identifiers such as Digital Object Identifier (DOI) names or Uniform Resource Names (URN) is a well known solution to the long term preservation of references. This approach is already widely used in long term preservation and the traditional publication world. For data access via the Internet, references provided by means of identifiers provide the location of the desired dataset in a way that is reliable and available over a long time ([PAS04]).

A persistent identifier clearly identifies units of intellectual property in a digital environment and serves for administration of these units irrespectively of form and granulation. It allows the citation of the digital resource (in our case dataset) and more importantly, identifiers allow also cross-linkage of digital resources, for instance, datasets to reference articles or to source datasets from which they have been derived. Finally, since the provision of the dataset identifier is achieved through a registration mechanism, it gives specialized actors of data curation the possibility keep track of the resource, index it in large catalogues and thereby dramatically improve the potential impact of a dataset publication.

All these aspects have been identified by the scientific community as valuable and crucial for a better usage of scientific datasets ([KLU06]).

A persistent identifier scheme always addresses two issues: The definition of the structure and syntax of the identifier itself; and the provision of a technical infrastructure for resolving. Today there are many different persistent identifier schemes used worldwide. The most common are URN, ARK, PURL and DOI.

URN: The formal description of the Uniform Resource Name (URN) was presented in 1994, its syntax was fully specified in 1997 as a standard from the Internet Engineering Task Force (IETF). There is, however, no central institution organising the URN; there is no central resolution infrastructure. The URN is more a general concept with isolated implementations. In 1999 the Conference of Directors of National Libraries (CDNL) introduced the National Bibliography Number (NBN) as part of the URN system. The major national libraries in Europe assign URNs starting with urn:nbn and offer a mutual resolving infrastructure.

There is however no central resolving infrastructure. When resolving a URN, it is always crucial to know where to locate the appropriate resolving mechanisms. Furthermore, there is no standard definition of metadata schemes. There are no

licence costs involved for assigning URNs. Each URN registration agency however has to establish an assigning and a resolving infrastructure.

PURL: The Persistent Uniform Resource Locators (PURL) were introduced 1996 by the Online Computer Library Center, Inc. (OCLC). PURLs are based on the http redirect mechanism. They offer a minimalistic technical approach in including a resolver address in the URL of a resource with a central resolver at the OCLC.

ARK: The Archival Resource Key (ARK) was introduced in 1995 by the California Digital Library (CDL). Like PURLs they are embedded in the htpp protocol and managed by the CDL as central organisation with central resolver.

DOI: The Digital Object Identifier DOI was introduced in 1998 with the funding of the International DOI Foundation (IDF). It is a registered trademark and DOI names can only be assigned by official DOI registration agencies that are a member of IDF. There are a total of currently 8 Registration agencies worldwide. The DOI system is technically based on the non-commercial Handle system of the Corporation for National Research Initiatives (CNRI). Since 2006, there is an ISO working group (ISO WG 26324) involved in the standardisation of the DOI system.

Registration agencies are responsible for assigning identifiers. They each have their own commercial or non-commercial business model for supporting the associated costs. The DOI system itself is maintained and advanced by the IDF, itself controlled by its registration agency members. Using the Handle system, there is a central free worldwide resolving mechanism for DOI names. DOI names from any registration agency can be by default resolved worldwide in every handle server; DOI therefore are self-sufficient and their resolution does not depend on a single resolution server. A standard metadata kernel is defined for every DOI name. Assigning DOI names involves the payment of a license fee by the Registration agency but their resolution is free.

DOI has emerged as the most widely used standard for digital resources in the publication world. It is currently used by all major scientific publishers and societies (Elsevier, IEEE, ACM, Springer, Wolters Kluwer International Health & Science, New England Journal of Medicine, etc.). The registration for the publishing sector is centrally run by the independent DOI Registration agency CrossRef, which assigns DOI names for 2609 members in the publishing sector. It is also used by the

European Commission through its publication agency the Office of Publications of the European Community (OPOCE).

Technically all of these persistent identifier systems could be used to register scientific datasets. The advantage of the DOI system lies in the possibility to establish citable datasets that can be handled as unique, independent scientific objects and are accepted as reference items by the STM publishers. The DOI system is well established and already part of the consciousness of the scientific community.

3.2 Citability through DOI names

While the interoperable and long-term preservation of linkage in scientific publication has been largely achieved through DOI over the last 5 years, dataset publication has not reached a similar maturity level. As mentioned in the last sections, the issue of access to datasets has grown more and more important in the different European research areas, none of these approaches however has yet established a workflow or a functional infrastructure for data registration.

A promising approach to establish dataset citation using DOI names has been started by the Organisation for Economic Co-operation and Development (OECD) for their own datasets. All statistical datasets published by the OECD in their annual factbook can be cited using DOI names [GRE09].

In the academic sector, an established approach within Germany that is actively used by scientists is the Data Registration agency for scientific data at TIB. TIB is the German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics, its holdings comprise around 7.3 million volumes of books, microforms and CD-ROMs, as well as around 18,000 subscriptions to general periodicals and specialist journals. TIB ranks as one of the world's largest specialist libraries, and one of the most efficient document suppliers in its subject areas.

In cooperation with several World Data Centers, over 600,000 datasets have been registered with DOI names as persistent identifiers by TIB. A selection of more than 1,500 datasets that are a part of scientific publications are furthermore directly accessible through the online catalogue of TIB and the German Common Library Network (GBV) ([BRA04]).

As a major advantage the usage of the DOI system for registration permits the scientists and the publishers to use the same syntax and technical infrastructure for the referencing of datasets that are already established for the referencing of articles. For example:

The dataset:

> Lambert, F. et al; (2008): Dust record from the EPICA Dome C ice core, Antarctica, covering 0 to 800 kyr BP, doi:10.1594/PANGAEA.695995

is used and cited in the article:

> Lambert, F. et al; (2008): Dust-climate couplings over the past 800,000 years from the EPICA Dome C ice core, Nature, 452, 616-619, doi:10.1038/nature06763

The citation of the dataset and of the underlying article follows the same standards and is therefore easy to adapt by scientists [ALT07].

Persistent identifiers are different from and complementary to local identifiers used in repositories. Local identifiers are useful for domain-specific applications or for local database management reasons. They can be used to reference the resource externally, but their validity is limited in time since such reference assumes the digital resource will remain in its current repository and that the repository structure will not evolve. Both assumptions are systematically proven wrong in the long run. By contrast, persistent identifiers are associated with the resource and remain identical regardless of the resource location; they are the preferred means for identifying the resource outside of the scope of the local system. Very often, a resource would have both a persistent and a local or domain-specific identifier. A common practise consists in building the persistent identifier from the local one at the time of registration. For instance, a DOI could look like: DOI:10.1594/***some domain specific ID***.

3.3 State-of-the-art at the International Union of Crystallography

Crystallography is a discipline that has benefited for a very long time from a tradition of data deposit and validation. Long-established databases such as the Cambridge Structural Database (for small-molecule structures) and Inorganic Crystal Structure Database (for inorganics) have abstracted crystal and molecular structures from journal publications and provide checking and curation services, as well as powerful

search and analysis software. The Protein Data Bank accepts deposits of structures of biological macromolecules, often in advance of journal publication.

The International Union of Crystallography (IUCr) publishes a number of journals, all of which require deposition in the journal archive of the data sets describing the final crystal structures, and in most cases the primary experimental data, or *structure factors*, from which the structure model is derived. The self-consistency and numerical integrity of these data sets is checked as part of the peer review part of the publication process. The checking relies heavily on a software suite, *checkCIF/PLATON*, which is made freely available for the use of prospective authors prior to submission, and is also available to other journals who wish to check submitted crystal structures.

Data sets accompanying journal articles are made freely available from the IUCr journals web site as supplementary data files, and are considered integral components of the articles that they underpin. As such, they are assigned individual DOIs by the CrossRef registration agency. In the CrossRef metadata schema, their parent/child relationship with the journal article is expressly encoded.

Other crystallographic data providers also use the DOI mechanism. The Protein Data Bank assigns DOIs to each macromolecular structure deposited therein. The UK National Crystallography Service, based at Southampton University, provides access to data sets that it has processed through institutional repository software. These data sets (which may or may not subsequently be published in the conventional scientific literature) also receive DOIs.

The IUCr has also used some of the services built by CrossRef around DOIs to publish relationships between other components of publications, and the way in which this has been done may have relevance to the need to identify various subsets of slices of larger resources (such as databases or data sets).

The IUCr reference series *International Tables for Crystallography* has been published since 1983 by Reidel, subsequently absorbed into Kluwer and in turn Springer. The series currently comprises eight large volumes, covering a broad range of topics such as symmetry and X-ray structure determination, electron and neutron diffraction techniques, fundamentals of crystallography, physical and chemical crystal data, and data exchange standards for the subject.

Recently the IUCr has published an online edition of the *International Tables*, in collaboration with the SpringerLink hosting platform. However, despite its extensive holdings and links between journal articles and book chapters, the SpringerLink site has limitations in expressing the close links between chapters, tables and databases within a collection of reference works dedicated to a single area of science. The online edition of *International Tables* was designed to be easily navigable and to take full advantage of the cross-references and links between the contents of the volumes in the series. The best accommodation with the SpringerLink functionality involved hosting static PDF page images of individual chapters on the SpringerLink site, while providing links to corresponding content in HTML format on the IUCr site.

Every PDF chapter on SpringerLink has a hyperlink to the corresponding HTML chapter on the IUCr site, where the contents of the volumes can be browsed, and each component viewed as a separate HTML page (by chapter, section, subsection, table or figure, as the reader wishes). The IUCr site also includes chapter indexes, links to related chapters, search engines, dynamic graph generators, hyperlinked symmetry-group relationships, links to a symmetry database, as well as PDF representations of section, subsections, figures *etc*. (see Figure 3)

Perhaps the most used parts of the *Tables* are the chapters that characterise the symmetry properties of space groups – the possible three-dimensional packings of a crystal lattice. On the SpringerLink site, these appear only as a contents list (in PDF format) containing hyperlinks. Each hyperlink leads to content on the IUCr site for a single space group (or sometimes for one of the possible multiple settings of some space groups). There are links both to PDF files (each of only a few pages extent) or to HTML pages.

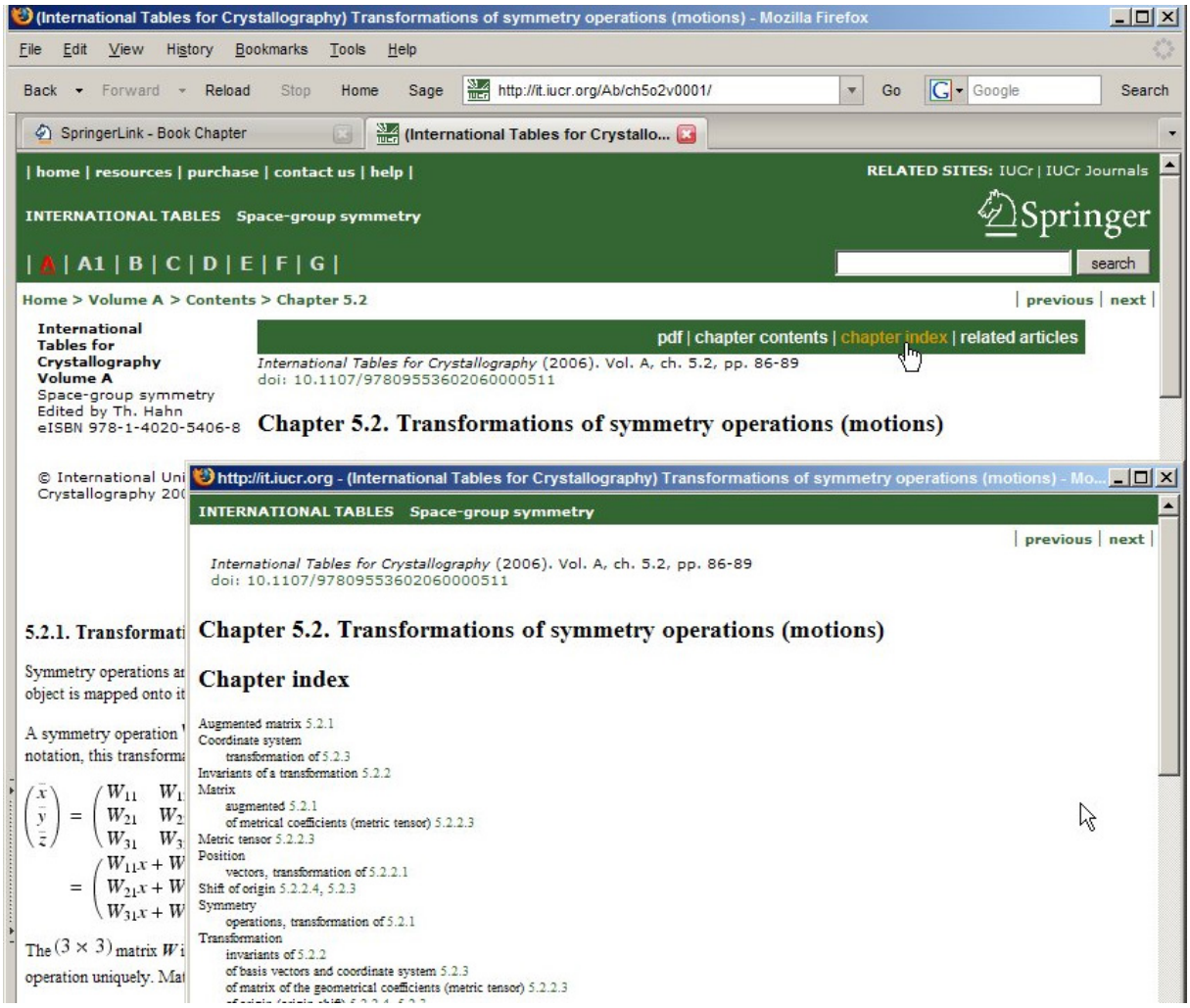

Figure 3 - Multiple hosting of content: SpringerLink site hosts PDF chapters where each PDF links to corresponding HTML pages on IUCr site. The IUCr pages have richer hyperlinking/other functionality than permitted by Springer XML schemata

The details of the linking have powerful ramifications. SpringerLink permits hyperlinks only to resources identified by a digital object identifier (DOI); and only a single DOI has been registered for each chapter. However, CrossRef supports a parameter-passing mechanism which allows an action to be associated with any DOI. We make use of this to link to a specific component of the chapter, via the chapter's deposited DOI.

In more detail, the hyperlink sends an openURL query to the CrossRef service. This query contains: a referrer identifier (useful if one wishes to tailor a different response to queries originating in different locations); the DOI, to allow identification of the primary resource (the chapter); and a payload, which carries a subordinate query that is modified by the CrossRef resolver into a call to another resolver at the IUCr. It is

the IUCr-resident resolver that finally transfers the user to the specific PDF or HTML file required.

The mechanism of global DOI registration provides resilient long-term associations between web resources. Adding to this openURL-style queries through a parameter-passing protocol opens the door to identifying and citing research data sets on an equivalent footing to literature citations. Furthermore, it opens the way to powerful query-language access to components of data sets. At a time when the association between scientific literature and supporting data is becoming ever more important, such techniques could revolutionise the publishing of data.

3.4 The Model of Data Registration at TIB

Since 2005, TIB has been an official DOI Registration Agency with a focus on the registration of research data. The role of TIB is that of the actual DOI registration and the storage of the relevant metadata of the dataset. The research data themselves are not stored at TIB. The registration always takes place in cooperation with data centers or other trustworthy institutions that are responsible for quality assurance, storage and accessibility of the research data and the creation of metadata. Figure 4 illustrates this structure in more detail.

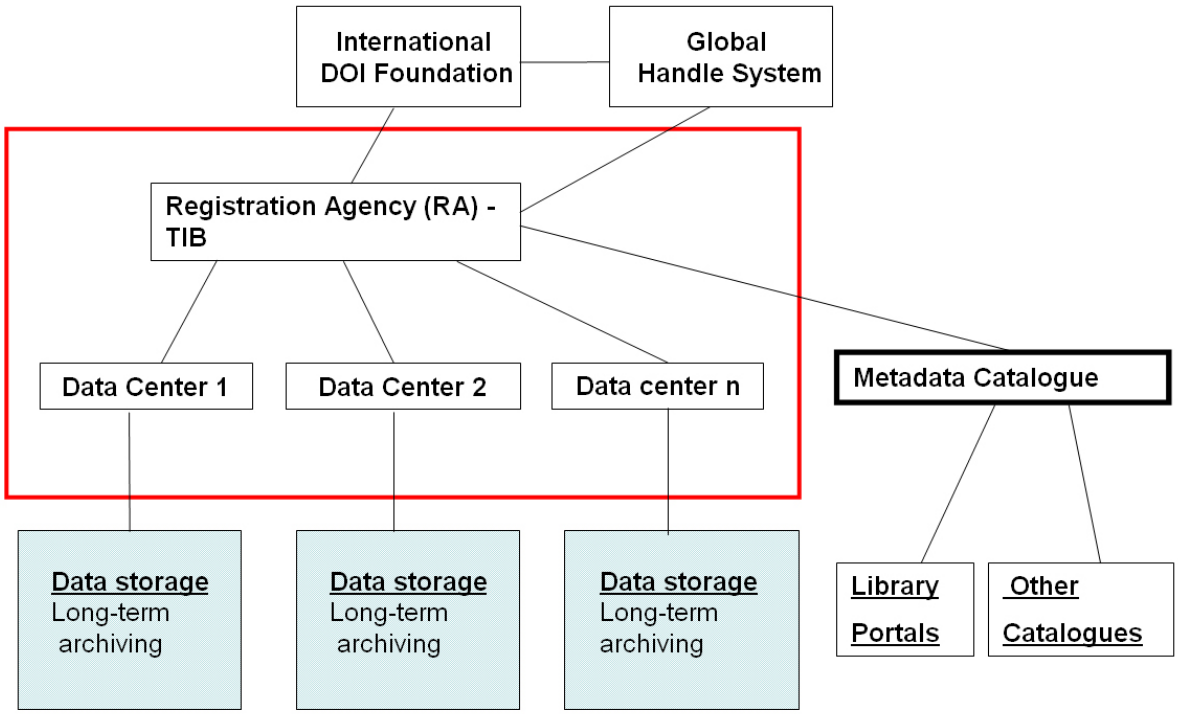

Figure 4 – *The overall structure of TIB's DOI Registration Agency*

Like for every persistent identifier, costs for infrastructure, personnel and license are involved for assignment of DOI names. TIB has three ways of re-financing its costs for the DOI license and infrastructure:

- TIB has customer-relations with data centres that receive DOI names for the content
- Costs for the registration of content that is of national interest are covered by the base funding of TIB as German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics.
- Registration of content that is a result of community funded research can be registered in cooperation with the funding agency by including the costs in the funding.

3.5 Dataset access through library catalogues

Library catalogues are classical sources for information [ING07]. The assignment of persistent identifiers allows further awareness of available datasets, when research data become directly accessible through library catalogues. When querying for a certain topic, users will not only receive all relevant publications as result, but also datasets collected by the corresponding researchers. Through dataset publication, researchers who collected data will gain further scientific reputation. This represents a further motivation for researchers to prepare collected data for online publication. Figure 5 shows the dataset mentioned above as result of a query to the online catalogue GetInfo of TIB.

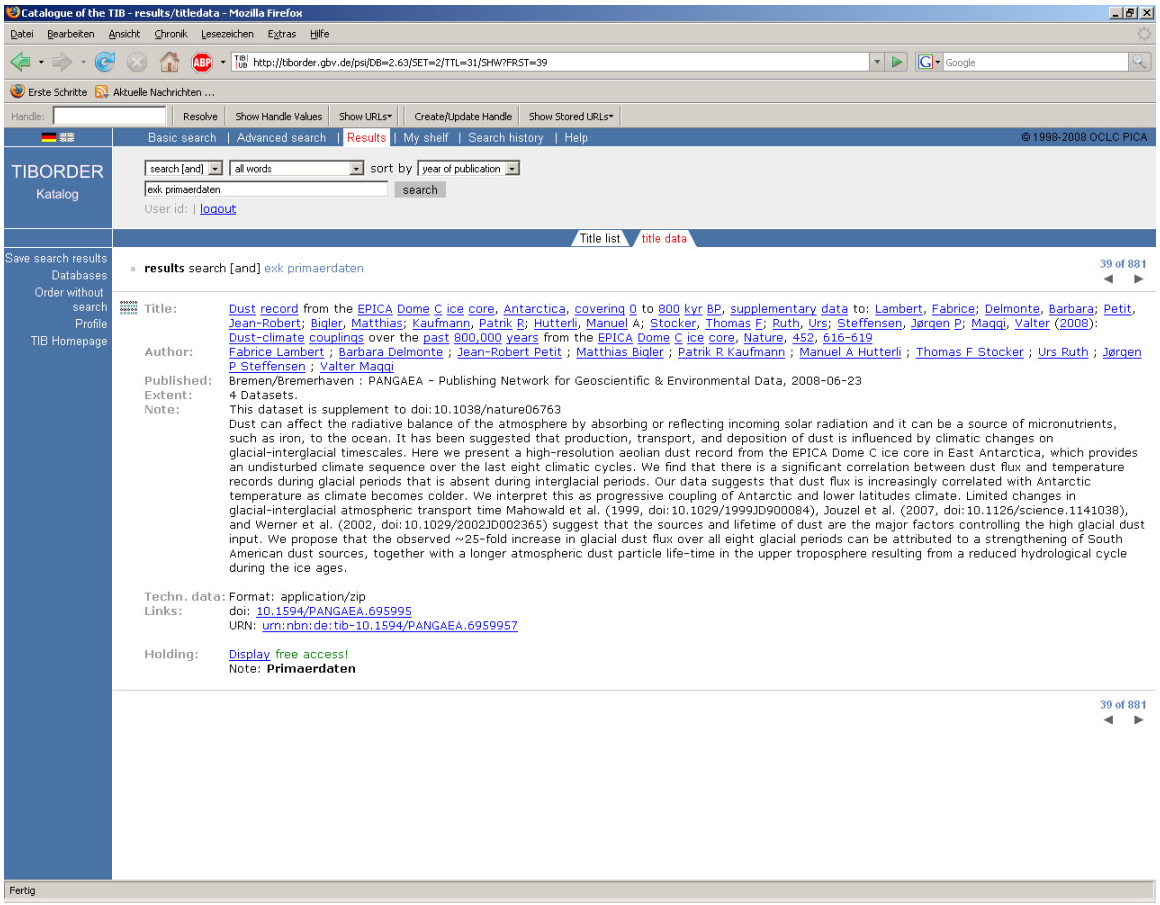

Figure 5- A scientific dataset as result of a query to TIB catalogue – GetInfo.

For a registered dataset the TIB stores all relevant bibliographic metadata about the dataset. This metadata is consistent with ISO 690-2 for the citing of electronic resources and is automatically mapped to the libraries catalogue format. There is however the need for more and better metadata schemes when dealing with scientific data. At present GetInfo is the only major library catalogue in Europe to include scientific datasets.

3.6 Dataset access through publisher pages

In a joint cooperation between TIB, the data publishing framework PANGAEA and Elsevier, scientific datasets will be accessible directly through the Science Direct page of the corresponding article. The workflow is based on a weekly harvesting of the PANGAEA data catalogue by Elsvier. Every DOI name of a dataset that is a supplement to an Elsevier journal article will automatically be included into the abstract page of this article in Science Direct as a link to “supplementary data” (see Figure 6)

Figure 6 – The Science Direct Page of an earth science article. The link "supplementary data" (highlighted) allows direct access to the underlying dataset hosted at PANGAEA.

4. Joint DOI Registration agency for scientific content

Access to research data is nowadays defined as part of the national responsibilities. As shown, during the last years most national science organisations have addressed the need to increase the awareness of and the accessibility to research data. Science itself nevertheless is international, scientists are involved in global unions and projects, they share their scientific information with colleagues all over the world, they use national information providers as well as foreign ones.

When facing the challenge of increasing access to research data, a possible approach should be a global cooperation for data access with national representatives.

- a **global** cooperation, because scientist work globally, scientific data are created and accessed globally.
- with n**ational representatives**, because most scientists are embedded in their national funding structures and research organisations .

The key point of this approach is the establishment of a Global DOI Registration agency for scientific content that will offer to all researchers dataset registration and cataloguing services. This joint agency shall be carried by non-commercial information institutions and libraries instead of publishers. This approach will allow easy access to the DOI system for non-commercial information institutes and libraries worldwide.

The objective of establishing an independent global DOI RA is to pool together resources of various interested local agencies. The benefits will be the following:

- Reduced infrastructure cost
- Better integration of the national infrastructures
- Reference implementation of the service in a distributed fashion
- Advanced distributed search capabilities for improving researchers' awareness of available datasets

Practical this new DOI RA can be implemented by widening the DOI model of TIB to a model of local agencies. This approach follows the example of the publishing industry in which the (often competing) publishers together use the central infrastructure of CrossRef to assign their DOI names.

Following TIB's model, data curation, maintenance and storage are not in the responsibility of the joint agency. Through its local partners it will furthermore offer services to existing national and international repositories and initiatives and therefore closing the gap between data infrastructure and information providers.

4.1 Roadmap

In a first phase the model of TIB will be opened to local agencies. These are libraries or information institutions with a national mission that includes the challenge of access to datasets. These local agencies will be direct partners of TIB and may use its infrastructure and license for DOI registration. On national level these local agencies will appear as directly responsible for the DOI registration (see Figure 7)

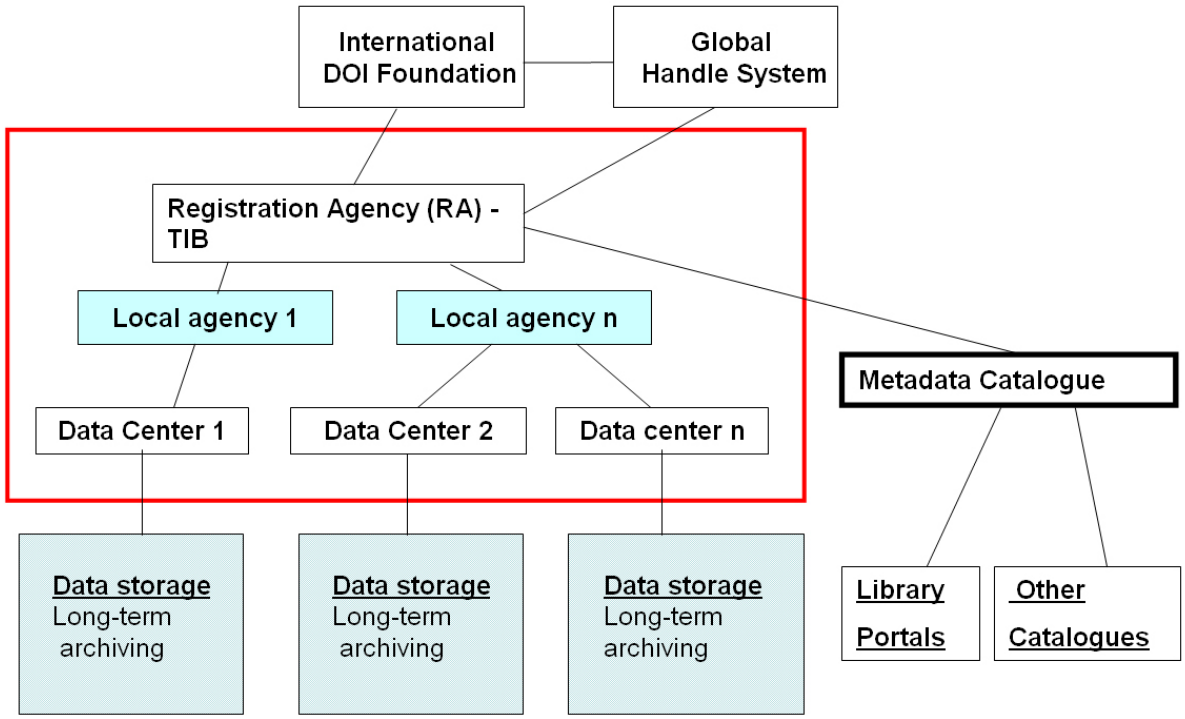

Figure 7 – The first phase of cooperation. National agencies as direct partner of TIB and responsible for their local data centers

In the second and final phase a new RA will be funded. This new RA will take the place of the TIB RA in the International DOI Foundation (IDF). It will be open for any information institute or library to join. The independent global DOI RA shall inherit TIB registration license and offer the existing services to other local institutions.

The structure of this DOI RA will be the following:

One central office will be located at TIB as the central address and responsible body for the International DOI Foundation (IDF), with a managing agent and technical staff. Each consortium partner will host its own office of the RA, allowing him to directly contact any data center in his domain. The partners are allowed to build up their own technical infrastructure for DOI registration or use the central infrastructure at TIB. If partners use their own handle server for registration these handle server will legally be operated by the joint RA. There will be one central metadata repository containing the descriptions of all registered data sets, with standardised interfaces to the partners own repositories and applications.

The metadata and workflow definitions will be standardised through all partners.

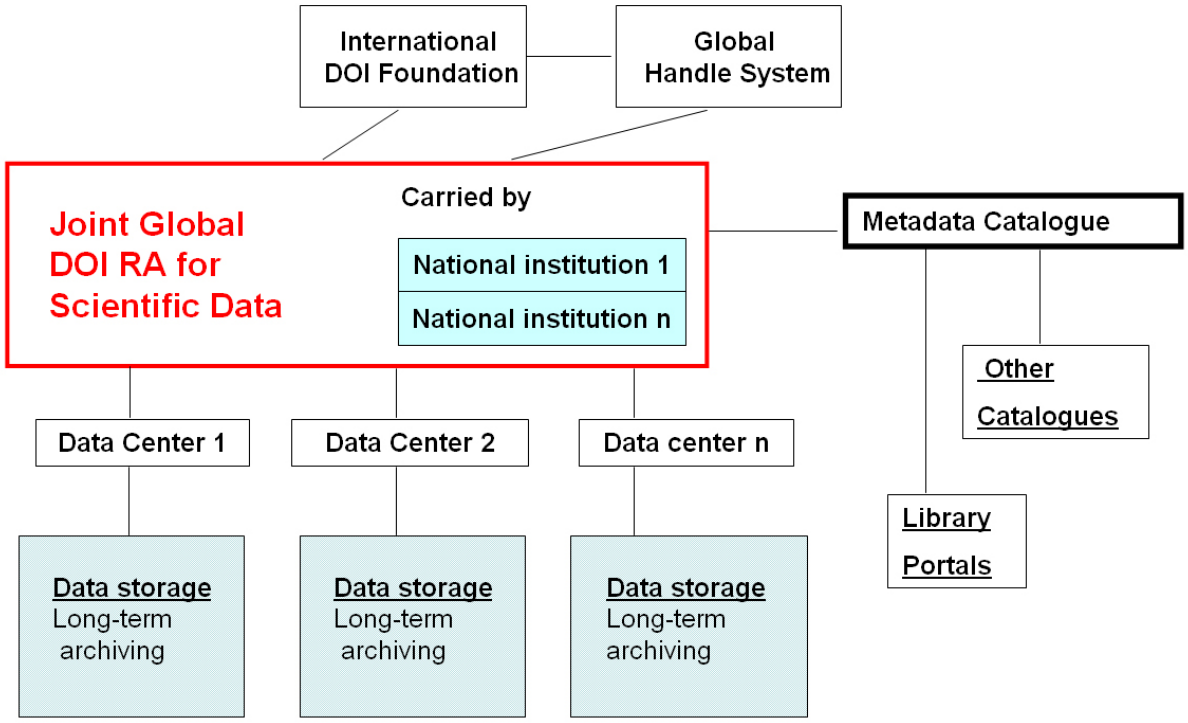

Figure 8 – In the final phase a new independent DOI RA take the place of the TIB RA.

Every partner including TIB will cover the personnel costs at their offices. The costs for the DOI licences and registered DOI names will be shared by all partners, weighted by the amounts of DOI names registered by each partner.

Every partner will have the right to develop its own business models for re-financing the registration costs.

The consortium will always remain open for other institutions to join under the same rules and obligations.

4.2 Partners

Institutions that have already expressed their interest to establish this agency are (in alphabetical order):

- **British Library (BL), UK:** The British Library (BL) is the national library of the United Kingdom. It is one of the world's largest research libraries, holding over 150 million items in all known languages and formats; As a legal deposit library, the BL receives copies of all books produced in the United Kingdom and the Republic of Ireland, including all foreign books distributed in the UK.
- **ETH Zurich Library, Switzerland:** The ETH-Bibliothek is the largest library in Switzerland and the main library of the Swiss Federal Institute of Technology.

In addition, it functions as the Swiss center for information on science and technology. The Library holds more than 6.9 million items, including maps, old prints, audiovisual materials, journals, databases and much more.

- **Institute for Scientific and Technical Information (INIST-CNRS), France:** INIST is a unit of the French National Center for Scientific Research (CNRS) under the administrative authority of the French Ministry in charge of scientific research. Its mission is to facilitate access to findings of all fields of worldwide scientific research. INIST-CNRS relies on one of the most important collections of scientific documents in Europe to provide a whole range of information services and Information portals providing access to electronic resources and dedicated to specific scientific communities.
- **National Technical Information Center Denmark:** The Technical Information Center of Denmark is DTU's center for scientific information provision, information management and information competences as well as the Danish national technical information center. The Technical Information Center of Denmark acts as a modern university library and as a center for management of the university's own research information. The information of the center is primarily disseminated and handled in a digital form and secondarily on the basis of printed collections. The public premises of the center are first and foremost designed to support the information searching and learning of the student.
- **TU Delft Library, Netherland:** TU Delft Library is the biggest technical-scientific library in the Netherlands. Its task is to safeguard the provision of technical-scientific information in the Netherlands. It focuses as much as possible on digital service in the field of technical science information. The TU Delft Library is the hub of knowledge for technical and scientific information in the Netherlands. It supports research and education within TU Delft and at the national level. The **3TU.Datacentre** is an initiative of the libraries of TU Delft, TU Eindhoven and the University of Twente under the auspices of the 3TU.Federation. The 3TU.Datacentre will provide storage of and continuing access to technical-science study data.

4.3 Memorandum

On 2 March 2009 the partners signed the following Memorandum of Understanding during the meeting of the International Council for Scientific and Technical Information (ICSTI) to establish a partnership to improve access to research data on the internet.

Memorandum of Understanding
Recognizing the importance of research datasets as the foundation of knowledge and sharing a common commitment to promote and establish persistent access to such datasets, we, the signed parties, hereby express our interest to work together to promote global access to research data.

Our long term vision is to support researchers by providing methods for them to locate, identify, and cite research datasets with confidence.

In order to achieve this long term vision, we will establish a not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers to them. The agency will take global leadership for promoting the use of persistent identifiers for datasets, to satisfy needs of scientists. It will, through its members, establish and promote common methods, best practices, and guidance. The organisations will independently work with data centres and other holders of research data sets in their own domains.

As a first step, this agency will build on the approach developed by the German National Library of Science and Technology (TIB) and promote the use of Digital Object Identifiers (DOI) for datasets.

Signed this day of March 2nd, Paris, France

Uwe Rosemann, Director, German National Library of Science and Technology, Germany
Wolfram Neubauer, Director, ETH Library Zürich, Switzerland
Herbert Gruttemeier, Head of International Relations, Institute for Scientific and Technical Information, France
Adam Farquhar, Head of Digital Library Technology, The British Library, UK
Mogens Sandfaer, Director, Technical Information Center of Denmark
Maria Heijne, Director, TU Delft Library, The Netherlands

5. References

- [ALT07] Altman M., King G., *A Proposed Standard for the Scholarly Citation of Quantitative Data*, D-lib Magazine, March/April 2007, Vol 13 No.3/4
- [ARZ01] Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. (2004) *Promoting Access to Public Research Data for Scientific, Economic, and Social Development*. Data Science Journal 3, 135-152.
- [AUS07] *Towards the Australian Data Commons, a proposal for an Australian National Data Service*, The ANDS Technical Working Group, October 2007
- [BRA04] Brase, J. (2004) *Using Digital Library Techniques - Registration of Scientific Primary Data*. Lecture Notes in Computer Science 3232, 488-494.
- [DIT01] Dittert, N., Diepenbroek, M. & Grobe, H. (2001) *Scientific data must be made available to all*. Nature 414 (6862), 393. doi:10.1038/35106716.
- [GRE09] Green, T (2009), *We Need Publishing Standards for Datasets and Data Tables*, OECD Publishing White Paper, OECD Publishing. doi: 10.1787/603233448430
- [ESF08] *Shared Responsibilities in Sharing Research Data: Policies and Partnerships.Report of an ESF–DFG workshop, 21 September 2007.*
- [ING07] Inger S., Gardner T*, How Readers Navigate to Scholarly Content, 2008* http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf
- [KLU06] Klump, J. et al *Data publication in the Open Access initiative*, Data Science Journal, Volume 5 (2006). pp.79-83 ISSN: 1683-1470 DOI: 10.2481/dsj.5.79
- [KO04] Koehler, W. (2004) *A longitudinal study of Web pages continued: a report after six years*. Information Research 9 (2). http://informationr.net/ir/9-2/paper174.html
- [LAW01] Lawrence, S et al (2001) *Persistence of Web References in Scientific Research*. IEEE Computer 34 (2), 26-31. http://www.fravia.com/library/persistence-computer01.pdf
- [NAT07] *Agencies join forces to share data* Nature *446*, 354 (22 March 2007) | doi:10.1038/446354b
- [NSF07]. *Cyberinfrastructure Vision for 21st Century Discovery*. Arlington/VA: National Science Foundation (NSF), Cyberinfrastructure Council (CIC).

- [OECD07]. *OECD Principles and Guidelines for Access to Research Data from Public Funding* Organisation for economic Co-operation and development, OECD publishing.
- [PAS04] Paskin, N. (2004) *Digital Object Identifiers for scientific data sets*. 19th International CODATA Conference, Berlin, Germany